Laboratories for Population Statistics
Manual Series No. 6
May 1975

Data Processing and Analysis in Demographic Surveys

Le Programme International des Laboratoires pour les Statistiques de la Population se consacre au perfectionnement des méthodes de mesure de variation des populations et à l'analyse des données démographiques. C'est dans ce but que le Programme cherche à établir et à subventionner des laboratoires d'expériences où de nouvelles méthodes de mesure puissent être expérimentées, et où, dans les cas appropriés, l'analyse de données déjà existantes puisse être développée. Ces projets sont basés sur la coopération entre l'Université de la Caroline du Nord à Chapel Hill et les institutions scientifiques de divers pays. Le programme se réalise grâce à la participation de ces institutions et à l'assistance apportée par l'Agence pour le Développement International du Gouvernement des Etats-Unis. Des publications scientifiques et techniques sont distribuées gratuitement à un nombre limité d'institutions et de particuliers qui s'intéressent aux problèmes de mesure démographique. Les opinions émises et les conclusions exprimées sont celles de leurs auteurs respectifs, sans engager les institutions dont ils dépendent.

El Programa Internacional para Estadísticas de Población se dedica a la
investigación de mejores métodos para la medición de cambios poblacionales
y al análisis de datos demográficos. Con este fin, el Programa espera establecer
y mantener laboratorios experimentales en los cuales se puedan ensayar
uevos métodos de medición y, cuando sea apropiado, analizar datos existen-
es. Los proyectos se basan en acuerdos cooperativos entre la Universidad de
Carolina del Norte en Chapel Hill e instituciones científicas de varios países.
El Programa se lleva a cabo gracias a la participación de estas instituciones
a la asistencia de la Agencia para el Desarrollo Internacional del Gobierno
le los Estados Unidos. Publicaciones científicas y técnicas son distribuídas,
in costo alguno, a un número limitado de instituciones y personas intere-
adas en problemas de medición demográfica. Las diferentes opiniones y
onclusiones expresadas son responsabilidad de los autores respectivos, y no
eben ser atribuídas a las instituciones patrocinadoras.

The International Program of Laboratories for Population Statistics is dedicated to research on improved methods for the measurement of population change and to the analysis of demographic data. Toward this goal the Program seeks to establish and support experimental laboratories in which new measurement methodology can be tested and, where appropriate, the analysis of existing data can be developed. Projects are based on cooperative agreements between the University of North Carolina at Chapel Hill and scientific institutions in various countries. This program is made possible through the participation of these institutions and the assistance of the Agency for International Development of the Government of the United States. Scientific and technical publications are distributed without charge to a limited number of institutions and individuals concerned with population measurement problems. The opinions and conclusions expressed are those of the respective authors and should not be attributed to the sponsoring institutions.

Laboratories for Population Statistics
Manual Series No. 6
May 1975

Data Processing and Analysis in Demographic Surveys

by Heather Booth and Joan W. Lingner

International Program of Laboratories for Population Statistics
Forrest E. Linder, Director

The Department of Biostatistics
School of Public Health

The Carolina Population Center

The University of North Carolina at Chapel Hill
Chapel Hill, N. C. 27514
U.S.A.

Contents

I. The Importance of Planning for Data Processing and Analysis

Less intellectually demanding than the formulation of the research design and less socially stimulating than the process of data collection, data processing and analysis is nevertheless a crucial phase of any research operation. The ultimate purpose of research is to provide answers or partial answers to the questions posed by the researcher and to make these answers available to policy makers, planners, scientists, and others. In particular, it should be noted that the research task is not ended when the investigators' curiosity has been satisfied.

Despite the fact that textbooks on the research process emphasize the need for early development of plans for analysis, this advice is seldom followed in practice. There are several reasons for its neglect. First, tabulation and analysis are, in a general way, usually implicit in the overall research design. Hence it may seem unnecessary to specify plans more precisely. Second, compared with the logistical problems of field operations, the frustrations of multiple callbacks, refusals, and the like, the processing and analysis of data may seem a relatively simple task. Third, because the data processing and analysis phases come at a relatively late stage in the research process, investigators may be strongly tempted to postpone planning these activities when faced with the myriad of decisions to be made concerning the research process occurring earlier in time.

Although the short shrift given to processing, tabulation, and analysis may be understandable, it is unfortunate. Careful planning for the analytic reports that are the end product of the research process as well as for each intermediate processing step may, if carried out at an early stage, clarify and sharpen the research design. In addition, careful consideration of the types of analysis to be carried out may assist in the elimination of questionnaire items of limited analytic value, and may also suggest the need for additional questions.

In this report, the steps in planning for the processing and analysis of data will be outlined and general guidelines for carrying out each step will be

HEATHER BOOTH *is a research assistant at POPLAB.*

JOAN W. LINGNER, Ph.D., *is an assistant professor, Department of Biostatistics, University of North Carolina at Chapel Hill, and is on the POPLAB staff.*

given. An important point to note is that, while the actual implementation sequence proceeds from editing the data through coding, processing, tabulation, and analysis to report writing, the planning of each phase is logically carried out in a somewhat different order. Since this report is intended for those responsible for overall coordination of research activities, the focus will be on planning rather than on implementation. However, it is hoped that the report will also be useful to those who carry out data processing and analysis.

The time sequence of steps in the processing and analysis of data in relation to other research activities is shown in Figure 1. This diagram exemplifies the use of PERT (Program Evaluation and Review Technique) in depicting the flow of research activities over time. Although further consideration of PERT is outside the scope of this manual, methods of constructing such charts and their uses in planning are discussed by Cook (1966) and by the PERT Orientation and Training Center (n.d.). In the present example, the abscissa represents time order, while lines represent activities and circles stand for events or completed activities. The number shown on each line indicates the expected duration of the specific activity, and the number in each circle stands for the cumulative time to the completion of an activity, and therefore indicates latest starting time for the subsequent activity. Dotted lines are used to link events that do not require time; lines radiating from a given point indicate activities that can be carried on concurrently. The *critical path* is the series of connecting lines within the network that represent the longest time elapsed from the start to the end of the project. The activities on this critical path are crucial in the sense that their completion will determine the total time elapsed for the project. Hence PERT charts serve as a tool for improved management. The development of a PERT chart requires that synchronization of all phases of the research operation be thought through carefully; such planning can be useful insurance against the processing and analysis phase being "lost" in the welter of other priorities.

In the present PERT example, rough tabulation and analysis plans are drawn up concurrently with the specification of concepts and variables. By way of an example, suppose that a national statistical agency wishes to undertake a survey with the following research objectives:

1. Estimation of current crude birth rates, age- and parity-specific fertility rates, and marital age-specific fertility rates
2. Measurement of cumulative fertility by age of women
3. Measurement of educational and urban-rural differentials in current and cumulative fertility and in age at marriage
4. Estimation of the family size expectations and aspirations of currently married women by age and parity
5. Estimation of the extent of knowledge about, attitude toward, and practice of contraceptive methods (KAP variables)

Given these objectives, the following groups of variables can be listed:

1. The total population
   - age
   - sex
   - current residence (urban-rural)
2. Women of childbearing age
   - current marital status
   - educational attainment
   - age at first marriage of ever-married women
   - number of marriages for each ever-married woman
3. Fertility of women of childbearing age
   - number of live births occurring during past year
   - number of children currently living at home
   - number of children currently living away from home
   - number of live-born children not surviving
   - total number of live-born children
   - current pregnancy status
   - number of additional births expected (including current pregnancy)
   - number of additional births wanted (including current pregnancy)
4. KAP characteristics of women of childbearing age
   - knowledge about methods of contraception by type of method
   - approval-disapproval of contraception
   - number and type of contraceptive methods ever used
   - number and type of contraceptive methods currently used

This preliminary specification of objectives and variables will serve to raise a number of questions about the research design, including sampling plans and questionnaire construction. For example, the sample must be selected so as to provide an estimate of the denominators for both crude and specific birth rates. The listing of variables also serves to point out the need for their precise definition. The lack of clear conceptualization and definition is especially obvious in the case of the knowledge and attitude questions, but also appears in more concrete variables such as educational attainment or live births. Careful definition of each variable will lead to the number and type of questions to be used in measuring it; this in turn will be useful in estimating total questionnaire length.

Finally, the list of variables can be used in preparing a preliminary set of tabulation plans relating to each research objective. Thus, for example, for the first objective the basic tabulations might be listed as follows:

List of Tabulations for Estimating Current Fertility Levels

1. Numbers of women of childbearing age by five-year age groups, annual numbers of births by birth order and age group of mother, and annual age- and parity-specific birth rates
2. Numbers of women of childbearing age by five-year age groups and marital status
3. Numbers of currently married women of childbearing age by five-year age groups, annual number of births to currently married women by age group of mother, and age-specific marital birth rates

This basic set of tabulations could be repeated for urban and rural residence and for each educational attainment category.

These preliminary tabulation plans have a number of important uses. First, a review of tabulation plans in the light of the research objectives will ensure that these objectives will be fulfilled by the analysis. Second, the specification of tabulation plans together with a general estimate of the number of categories to be presented for each variable will permit estimation of the total number of cells required for each tabulation. Thus, for example, if seven quinquennial age groups are used for women aged 15 to 49 and birth order is classified into 12 categories (0 to 11+) the number of cells required for the distribution of births by parity and age of women is 7 × 12 = 84. This estimate of cell requirements will be of value in designing efficient table formats, as discussed in Section VII. Third, the tabulation plan will serve as a logical starting point for a careful inventory of data processing equipment and its capabilities as discussed in detail in Section VI of this report.

Each subsequent step of the data processing and analysis operation will be related to the start or completion of other aspects of the research process. Thus, little can be done toward the development of coding frames until a preliminary questionnaire has been drafted; on the other hand, if the questionnaire is to be precoded, the coding frames should be completed before the final draft of the questionnaire is prepared. Similarly, editing routines will normally be based on experience gained from pretests and pilot studies and cannot be put into final form until such studies are completed. The availability or nonavailability of computers, packaged programs, and the like will also have an important bearing on coding and editing strategies. Care taken in the early planning stages to ensure that the data collection and coding methods are compatible with the processing plans will avoid delays resulting from recoding of cards or other transformation operations.

Subsequent sections of this report will discuss in detail the construction of coding frames; the development of editing procedures; coding, punching, and data storage procedures; and the preparation of tabulations. Thus the order of presentation approximates the order followed in planning these operations. It should be reemphasized that the planning of these tasks requires that each aspect be neatly dovetailed with every other, and hence many planning activities must be concurrent.

II. Construction of Coding Frames and Scales

A. INTRODUCTION

As noted in the preceding section, the first step in planning the data processing operation is the specification of research objectives, variables, and preliminary tabulation plans. Codes should be constructed in accord with these specifications; this should be done by the researcher rather than by the clerical staff since it requires an overall knowledge of the study and its aims. An equally important subsequent step is the review of the available mechanical and electronic data processing equipment. (A glossary of relevant computer terms appears in the Appendix.) For the purposes of this manual, it is assumed that data processing equipment is available and, further, that it uses the conventional 80-column punch card as one of the possible means of input. Various specific features of the punch cards will be relevant to the design of coding frames; these are discussed in Section II-D below.

B. SOME GENERAL CONSIDERATIONS

A *code* is an abbreviated and simple means of describing and recording information. Because of processing equipment requirements, codes are usually numeric; thus the numeral 1 may be used as a code for male, and 2 for female. The set of categorized possible responses to a given item is referred to as the *coding frame*; the set of coding frames covering all items on a questionnaire is known as the *code book*. The order of the coding frames within the code book should be the same as that of the questions on the questionnaire. The code book should also contain information on the location of the data on the punch card. Sample pages from a typical code book appear in Figure 2.

The main purpose of using codes in data processing is to enable the data to be handled and processed by the processing machinery. However, the construction of coding frames must reflect the underlying aims of research. The first step in the process of code construction should be the development of the concepts, definitions, and categories of interest to the research study. Once formulated, precise definitions of the variables corresponding to each concept should be adhered to throughout all stages of the work. It is important, therefore, that the questionnaire be designed and the data be categorized and coded in accord with these definitions. Extracts from a fertility questionnaire corresponding to the coding frames in Figure 2 are shown in Figure 3.

Where possible, definitions and concepts already in existence (such as the recommendations contained in the United Nations *Handbook of Vital Statistics Methods* [1955]) should be used, especially where comparison with previous data is to be made. Use of existing coding frames is also work saving. In addition, coding frames should be tested whenever possible on samples of replies; this step is essential in developing codes for previously untried questions. Use of samples of actual responses also aids in the detection and elimination of ambiguities and troublesome codes.

When previously untried questions are to be asked it is of course necessary to devise appropriate coding frames. The number of categories for each question depends on the level of detail required in the analysis; while they should be few enough to summarize the data effectively, it is advisable to retain more detail rather than less since it is possible to amalgamate but not to split groups after the coding has been completed.

Understandably, simple questions are easier to categorize than complex questions. Hence, it may be better to ask several simple questions of the respondent than to ask one complex question: responses to these questions can be coded individually. Subsequently, the simple codes can be combined to form more complex ones.

C. OPEN, CLOSED, AND PRECODED QUESTIONS

Certain questions, particularly those pertaining to opinions and attitudes, can be asked without giving the respondent any suggestion as to possible replies. Responses to such questions are copied verbatim by the interviewer and coded after the interview has been completed. These questions, which are sometimes termed "open" questions, require the development of coding frames that permit the wide variety of responses elicited to be classified into relatively few categories. Appropriate coding frames may be built up from responses obtained during the pretest of the questionnaire. Two considerations are particularly important in developing codes for open questions. First, the categories must be sufficiently unambiguous and nonoverlapping that responses can be classified consistently. Second, although an "other" category for responses that do not fit elsewhere in the coding scheme is usually necessary, the number of responses actually coded "other" should be small. If results from pretesting indicate that these two conditions cannot be met, it may be advisable to transform the question to a "closed" question which forces respondents to choose among alternative prespecified responses. The alternatives can be presented by reading the choices aloud or handing the respondent a checklist of permissible answers. Question 41 of Figure 3 is an example of a closed question of this type, while Question 41B is an open question. Questions on such characteristics as age, sex, marital status, and number of children living at home usually give rise to quite predictable responses and can also be treated as closed questions.

Closed questions are readily adapted to use in precoded questionnaires where the categories are marked on the questionnaire and each has its code alongside. The interviewer then marks or circles the appropriate code. Incorporating the coding process into the interview schedule thus obviates the need to code at a later stage. For example, the question "What is the sex of the child?" can be answered by either "male" or "female." These two possible responses may be coded by designating male to code 1 and female to 2. Both the category descriptions (e.g., female) and the codes appear on the form. If a respondent answers male, the interviewer circles or marks code 1; thus there is no need to code this question later on.

It may be necessary to include an "other" category to allow the use of precoding; this may be done with or without a request for detail which may be of sufficient interest to be classified and coded at a later stage. The list of contraceptive methods shown in Figure 3 includes an example of a precoded "other" category.

An additional feature of precoded questionnaires, also shown in Figure 3, is that they include the appropriate column number indicating the location of the datum on the punch card. It is then possible for the data to be punched directly from the information recorded on the questionnaire at the interview stage.

D. TYPES AND PROPERTIES OF CODES

Since it is assumed that standard 80-column punch cards[1] are to be used in data processing, codes must be developed within the restrictions of card capacity and processing requirements. The first (or last) several columns of each card must be reserved for identification purposes, which reduces the number of columns available for recording data on other variables. Typically the data for each unit of analysis (households, individuals) are recorded on one or more punch cards. If more than one card per case is required, each card must contain both the case identity number and a card identification number. In surveys where the number of variables is small, it may be physically possible to record the data for more than one unit of analysis on a single card. Although this reduces the number of cards used, it is almost always inadvisable to do this. Problems arise at the tabulation and analysis stage, since cards will have to be processed more than once. The saving involved in card costs will usually be more than paid for in inconvenience.

The 12 punch positions per column are subdivided into two sets. Numbers 0 to 9 occupy the lower 10 positions, while the upper two punch positions (zone positions) are variously referred to as X and Y, 10 and 11, X and V, – and +, and L and U for the lower and upper respectively. Although zone positions can be used to increase the capacity of a column, it is usually wise to restrict numeric data to the lower 10 punch positions, since both the hardware and software used in analysis are often unable to interpret zone punches. Use of zone punches or blanks to represent codes may therefore lead to expensive programming to recode the data into an acceptable form. Further, the use of multiple punches in a single column weakens the card and increases the possibility of machine jams. Alternative approaches to economizing on the number of columns used through multicharacteristic codes and geometric codes are discussed below.

Coding categories should be both mutually exclusive and exhaustive so that no ambiguities arise and every possibility is covered. For example, the variable "age in completed years" can be categorized into, say, single year of age groups 0, 1, . . . 95+. It is clear that each age must fit into one and only one age code. It should be noted here that the final category, 95+, is open-ended to allow for all reported ages of 95 and above.

Provision should also be made for coding nonresponses to specific items, including cases where the requested information is not known, is refused, or is missing, as well as cases where the question is not applicable to a particular respondent. In general the "not applicable" category should be coded separately since this affects the base number of item responses used in calculations. The "not known," "refused," and "missing" categories may be coded together as one category unless identification of these types of nonresponse are of separate interest (as they may be for purposes of quality control). In that case a breakdown into the three separate categories may be desirable. It is usually beneficial to reserve the same code for nonresponses throughout the questionnaire. Thus codes such as 9, 99, 999, and so on, can be used to denote nonresponse. Further, it may be useful to reserve 8 and 98 for "not applicables." This greatly increases the ease with which such codes can be identified throughout the whole data set and also allows for some or all of them to be systematically omitted from the data for purposes of statistical calculations. See Figure 2 for examples.

A further requirement for a good coding frame is that it should provide a logical framework for viewing responses. Categorization of responses implies the construction of a scale. In general, there are four scale types in which the data may be coded.

Nominal scales classify the responses into two or more groups without implication of rank or distance between groups. This scale applies to nonquantitative information such as sex and marital status.

The second type of scale is the *ordinal* scale where the groups are ordered or ranked along a characteristic scale but there is no implication of distance between the categories. A variable suited to this scale is birth order of children where the code number is identical to the rank.

A more sophisticated scale is the *interval* (or *cardinal*) scale which has equal units of measurement enabling the distance between ordered categories to

[1] Other kinds of cards exist but they require the adaption of standard machinery or even use of special machinery. Before these cards are used, the researcher must be absolutely certain that he will not require further analysis which is possible only on standard machinery. Punched paper tape is also occasionally used with some types of data processing equipment. At best, however, it serves as a very temporary form of storage, since the tape is quite fragile.

be interpreted. This scale is additive but is not multiplicative since the zero position is arbitrary: the distance between -1 and $+1$ equals that between 3 and 5, but 10 cannot be regarded as twice as much as 5. (A good example of this kind of scale is a thermometer.)

The highest level of measurement is the *ratio* scale which has all the properties of the interval scale with the addition of a fixed zero. Both differences and relative magnitudes can thus be compared. Examples of this kind of scale include weight and age.

In categorizing the data, both the distribution of the variable and the research objectives should be considered. Where there would be very few cases in a category it is often practical to combine adjacent categories. This is particularly relevant where a variable distribution has rather a long tail. The final category probably should be open-ended and may in fact be several times as wide as other categories of the variable. In such cases category width must be taken into account in the estimation of certain statistics such as means and variances.

Whatever the type of scale employed, codes are usually constructed by using consecutive numbers. However, situations may arise in which nonconsecutive number codes are advantageous. For example, where a variable is represented by a set of nonconsecutive numbers (such as day and month of birth) it may be useful to retain the original numbering system as codes. Another instance in which nonconsecutive codes may be useful is in the construction of branch codes which accommodate use of the data at several levels of detail. The basic principle underlying this type of code is to split the variable into several broad categories which are each further divided into subgroups, and so on. Thus, 041235 might be used to signify the thirty-fifth sub-subgroup of the twelfth subgroup of the fourth group. For example, in a polygamous society children may be recorded by birth order and by mother. The first digit would then contain the broader information (i.e., mother) and the second and third digits the detailed information (i.e., birth order to that mother). Either level of detail can then be used in the analysis. The resulting distribution will be discontinuous, since 112 representing the twelfth child of the first wife may be followed by 201 representing the first child of the second wife. This type of coding may be used to represent geographical divisions where the country is divided into provinces, the provinces into districts, and the districts into villages; information is then readily available for each level. In general, branch coding requires more digits (and therefore more space on punch cards) than a straightforward sequence code. For this reason it should be used only when different levels of detail are required in the analysis.

Decimal codes are a special case of branch codes in that there are no more than ten divisions at each breakdown stage. Customarily the codes are written with a period or full-stop after each group of three digits to facilitate reading. It should be noted that the use of decimal codes is advantageous only where large numbers of categories are to be coded, since it requires greater effort and time than simple sequential coding and increases the risk of error in coding, punching, and tabulating.

Occasionally column economy can be gained by the use of multicharacteristic codes. For example, sex and marital status require two and four punch positions respectively, each variable requiring one column. Since these variables are often referred to in conjunction (e.g., never-married females), it may be advantageous to code the eight possibilities in one column: that is, never married, married, separated and divorced, and widowed are coded as 1 to 4 for males and 5 to 8 for females. In the event of all males or all females being required, the categories can be collapsed. It should be noted that information will be lost by the use of such codes because missing information on either sex or marital status results in an unknown code of 9.

Geometric codes may be used in the case of multiple responses where the respondent may fit into several categories simultaneously. For example, several methods of birth control may be used by a respondent; if the checklist contains five methods, the first should be assigned a value of 1, the second 2, the third 4 ($=2^2$), the fourth 8 ($=2^3$) and the fifth 16 ($=2^4$), (and in general the *nth* should be coded 2^{n-1}). Then the code for a particular respondent is the sum of the assigned values of the methods that are used. Thus, for example, a response indicating use of both the first and third methods is coded as 5 ($1+2^2$). By using this system, each combination of methods is uniquely represented by one code and the number of columns necessary to accommodate the codes is only two. If each method had been coded on a used or not used basis, five columns would have been required. (For a further example see Figure 3.) Simple frequency counts can be obtained from these codes; an example of this is shown in Figure 5-C.

It should be noted that more complicated types of codes such as geometric codes impose higher skill requirements on coders and editors. Thus the card economy gained by their use may be more than offset by the introduction of coding and editing errors or the need for additional training.

III. Editing Procedures

A. PURPOSES AND TYPES OF EDITING

The editing process is intended to detect and eliminate errors in the data. The most basic editing operation is to check for completeness: there should be an answer to every applicable question. Where replies are missing it is often possible to determine whether the missing information pertains to a "not applicable" question or is a nonresponse. "Not applicable" responses can be appropriately coded. The treatment of nonresponses depends on the type and stage of editing.

In addition to checking the completeness of the questionnaire, the accuracy of the responses must also be evaluated. This involves examining the consistency of each record. Consistency checks involve comparisons of related characteristics. For example, if an individual is reported as never married but year of first marriage is given as well as a non-zero response to number of marriages, the never-married response is obviously inconsistent. Types of checks to be made and the procedures to be followed when inconsistencies are found must be clearly specified.

B. STAGES IN THE EDITING PROCESS

Editing activities can take place at several stages of data collection and processing. The first opportunity for editing is in the field, and it will normally be undertaken by the field supervisor as interviews are completed. Checking for completeness at this early stage is important, since the chances of an interviewer's being able to clarify details and to supply or to call back for missing information is greater if questionnaires are edited promptly in the field. Thus, field editing should reduce the need for later field verification reinterviews and the associated costs.

Field editing can also play a useful role in the training and supervision of interviewers. Field editors should make certain that instructions have been interpreted uniformly by the interviewers. This is very important since the data are ruined if the interviewers fail to ask all applicable questions. If relatively inexperienced interviewers are employed, rigorous field editing is essential.

Given the pressures of fieldwork, however, it is unlikely that the editing carried out at this stage will be as complete and thorough as is desirable.

Accordingly, in most cases some provision for central office editing is needed. Depending on the size and complexity of the survey, this editing may either take place before coding the data or may be combined with the coding operation.

Since they are also able to read through an entire questionnaire to develop a "feel" for the proper interpretation of responses, central office editors can take advantage of the clues that may be provided in the stray comments and notes the interviewer may have included. (Too much exercise of intuition is not to be recommended, however). If missing information is not detected until questionnaires have been sent to the central office, it may be necessary to return the questionnaires to the field for checking. If this is not possible and there is no other way of obtaining the information, the question must be coded as a nonresponse. Values for nonresponses may be imputed at a later stage (see Section III-C).

After the data have been coded and punched, data processing equipment can be used for further editing. Two types of machine edits can be distinguished: range edits and consistency edits. Range edits are used to check whether each value assigned to a variable corresponds to a meaningful code. Thus, if the allowable codes for sex are: 1 for male, 2 for female, and 9 for no response, then the range edit should identify those records having codes of 0 or 3 through 8. The consistency of the data can also be checked by machine. For example, if both age and year of birth are asked, they can be checked against one another, marital status can be compared with age, and such unlikely combinations as widows aged 13 can be identified and checked. Machine edits are particularly valuable in detecting coding, editing, and keypunching errors made at earlier stages. As will be discussed in the next section, procedures can be developed to impute missing data at greater or lesser degrees of sophistication.

All three types of editing, field editing, manual central office editing, and machine editing, are likely to be useful in improving the quality of the data, but the optimal combination of these editing options depends on the size and complexity of the survey. It may be most efficient to focus the energies of field editors on checking the completeness of response, and to carry out more complex consistency checks at the central office.

C. IMPUTATION

An extremely important step in the preparation of data for analysis is the development of procedures for imputing values for missing or inconsistent responses. Simmons (1972, p. 41) has noted that:

> The first point to recognize is that when data are missing, imputation *must* take place, either implicitly or explicitly. If the results of the survey are to be used at all, the analyst or consumer is compelled to draw conclusions about the missing evidence. It is not a question of *whether* to impute, but *how*.
>
> Some analysts might hesitate to "fabricate" data for missing entries, thinking that to do so is cheating. But suppose for a given item, 10 per cent of the entries are missing. If no explicit action is taken, aggregate values from the survey will be of the general order of 10 per cent low, while consumers of the report will be asked to believe and indeed forced to assume that with respect to rates or ratios, the 10 per cent missing cases have the same average value as the overall average of the 90 per cent which were reported. Often this will be an unjustified assumption. Usually a superior imputation procedure is available.

Certain types of imputation can be built into the routines for checking completeness and consistency of response. For example, if both year of birth and current age are asked, but only one of these questions is answered, the editing procedure can be used to supply the missing datum on the basis of what is given. If this editing rule is adopted, then procedures for imputing the appropriate response must also be spelled out in detail. Thus, since a person born in 1930 might be either 42 or 43 in 1973, an unbiased procedure such as random assignment of date of birth must be developed. Other acceptable imputations can easily be devised: if no entry for marital status is given, but marriage duration and husband's current employment are reported, it is usually safe to assume the respondent is married.

Certain characteristics however cannot be imputed on the basis of other information in the record. In such cases, imputation is best carried out by assuming that the characteristic is similarly distributed for the respondents and nonrespondents in the same subclass. The problem of course is to determine the appropriate subclass. Simmons (1972, p. 42) suggests three rules for making the selection:

> A. Choose subclasses for which the difference . . . [in the distribution of the characteristic] . . . is, or is likely to be, substantial.

B. Choose subclasses for which non-response rates are different.

C. Choose subclasses which are of sufficient size . . . that they have more than trivial impact on overall results, and at the same time assure that there will be enough responses in the subclass to yield a stable estimate of the imputed values.

Age, sex, place of residence, and marital status may all be useful in defining subclasses.

For some analytic purposes it may be sufficient to carry out imputation at the tabulation stage by prorating the nonrespondents in each subclass over the distribution found for the respondents. Thus, if there are 100 respondents and 10 nonrespondents to a question on marital status in a particular age-sex group, the distribution of responses can be adjusted by multiplying the observed frequency of each marital status category by 1.10, i.e., (100 + 10)/100. In many cases, however, it may be desirable to replace nonresponses and erroneous or inconsistent responses in the individual records with the imputed values. This can be done by allocating the records of the nonrespondents in a defined subclass in proportion to the distribution observed for the respondents. The allocation procedure should ensure random assignment of records to categories within a subclass. It is important that a count of the number of imputations made be maintained. Variables for which there are high rates of nonresponse are, of course, to be analyzed with caution.

IV. Coding and Punching

A. METHODS OF RECORDING CODES

The actual coding of the data should be a relatively straightforward procedure if the requirements for good codes have been met adequately and the editing has been thorough. Where the questionnaire is simple and consists entirely of closed questions, the coding stage can be incorporated into the data collection stage by use of precoded questionnaires, as discussed in Section II-C. After editing, the punching operation can then be done directly from the questionnaires.

In cases where some or all of the questions are not precoded, these questions may be coded centrally by writing the appropriate code in a designated place on the questionnaire. Again, the punching is done directly from the questionnaires. These two methods can be readily intermixed, the simpler questions being precoded and the more complex being coded in the office.

Where punching is to be done directly from questionnaires, it is essential to the keypuncher that the column numbers corresponding to each code also appear on the questionnaire. (See, for example, Figure 2). In addition, keypunching will be facilitated if the codes are clearly written or circled, preferably in colored pencil.

An alternative method of recording codes consists of writing them out on a separate form called a *coding sheet*. This sheet is designed to be parallel to the punch cards, that is, each line of the sheet provides for 80 digits (or occasionally 40 or 120). Coding sheets must be used when the questionnaire is not designed to accommodate the codes. Their use may result in some saving of keypunch time in that many records can be punched before a page is turned. In general, however, the savings in coding time from the use of precoded questions and from writing codes directly on to the questionnaire will more than offset savings in keypunch time. Further, the transfer of data onto coding sheets may result in the introduction of copying error.

Perhaps the most efficient approach to coding and punching is the use of *mark sensing*. This is a method by which a machine reads pencil marks from questionnaires or coding sheets and punches holes appropriately. For precoded items, the pencil marks can be entered at the time of the interview, while the marks corresponding to responses to open questions can be entered by central office coders.

The method removes the human error involved in the punching process, though the error in the coding and interviewing process is still present. Recent innovations in computer technology permit the data to be punched directly onto computer tape, or a character reading device transfers the data automatically from the document to the tape. In the United States special equipment "reads" microfilmed copies of the census questionnaires and transfers data directly to computer tape (Fasteau et al., 1964).

B. TRAINING AND SUPERVISION OF EDITORS AND CODERS

The training program for editors and coders should embody a thorough explanation of the goals and purposes of the study and a close examination of the rationale for each question asked and its relation to the concepts and variables being studied. Training materials should include code books, coding and editing manuals, code sheets, and examples of completed questionnaires. It may be useful for editors and coders to observe actual interviews.

Coding and editing are hard work. An important task of the supervisor is to see that work circumstances are conducive to good work. This includes making the environment as agreeable as possible by providing adequate and pleasant work space, control of noise, and accessible supplies and equipment. It also implies consideration of human reactions to performing detailed tasks meticulously. For example, it may be advisable to instruct workers to take frequent, short breaks.

In the earliest stages of editing and coding—which may well be during the pretest or pilot study—supervision should be extremely close, since problems not anticipated in earlier planning are likely to be discovered. Editing procedures and coding frames may need to be modified on the basis of this experience. As workers become more familiar with their tasks, supervision can be less intense, but it should always include regular review of edited and coded questionnaires. In particular, it should be noted that the supervisor's role is more than that of a troubleshooter who participates in the editing and coding process only after problems arise, but rather it is one of ensuring that operations are carried out smoothly, accurately, and efficiently.

In addition to day-to-day supervision, attention should be given to more formal aspects of quality control. For example, one approach to the control of coding errors is to have records checked by a second coder or coding supervisor. However, the checker may be influenced by the original coder's judgment. A more rigorous approach would require independent coding by two or three coders and careful reconciliation of discrepancies (Minton, 1969). Since this approach is costly, its use is usually restricted to a sample of records. Independent coding may be especially useful when applied to pretests and pilot studies since common types of errors may be located and ambiguities in code books clarified. Human error is also involved at the punching stage and is checked by means of a machine known as a *verifier*. The punched cards are fed into this machine and the operator "repunches" the data though no holes are in fact made. As long as the punched holes and the verifying punches agree the card moves through the machine in the usual way. Where the two differ, a light comes on and the machine locks: the operator determines the correct punch and if necessary repunches the card. Verification is recorded by a small notch on an edge of the card. The chance of the same mistake being made twice leaves room for error, though the chances are small. However, where codes are illegible, the chance of this type of error is increased.

C. CONTROL OF RECORDS PROCESSED

At each stage in the process, checks should be kept on the exact status of each part of the data. From the time each interview is completed to the preparation of final reports (and even beyond if additional analyses are contemplated), carefully defined and tightly controlled procedures should govern the location of the data. Usually it will be convenient to process questionnaires in *batches* according to, for example, the period when the interview was completed or according to a geographic segment. After the field supervisor has reviewed the questionnaires from a given batch, they should be counted and bundled together (with rubber bands, tapes, or suitable clips). A control slip showing the number of records in the batch (and other identifying information) is then affixed to the bundle. When the batch is forwarded to the central office, its receipt and identification will be recorded on a batch control chart. The batch will remain together as a unit as it passes through further editing, coding, and punching phases; at each step the questionnaires should be counted to be certain that their number agrees with that on the control slip. If a record is found to be missing at any stage, steps should be taken to retrace its progress until it is located or, at a minimum, until it has been identified. Completion of each step should be noted on the batch control chart before the batch is passed along to the next stage. Hence, it should be possible to locate any given record at any time by reference to the batch control chart.

The number of cards punched per batch should of course be equal to the number of cards per questionnaire times the number of questionnaires. The number of cards per questionnaire may vary, as, for example, when the questionnaires contain data for all numbers of the household and a separate card is punched for each individual. In such situations it may be easiest to sort each batch of questionnaires by number of persons in the household and hence to calculate the number of cards to be punched. Both the number of cards and number of questionnaires processed should be entered on the batch control chart.

V. Storage of Data

A. ORIGINAL RECORDS

Even though all the information has been coded and punched onto data cards, it is still advisable to keep the original questionnaires at least until the analysis is completed. Though more than one copy of the data should be kept, it is possible for the punch cards or other data storage device to be destroyed or misplaced. In such cases, it may be necessary to repunch the information from the original questionnaires or from coding sheets if available. Further, original questionnaires are the only means of resolving any inconsistencies and errors in the data that become apparent during the analysis.

B. ENCODED RECORDS

As noted in the preceding section, it is always advisable, whatever the form of data storage, to keep more than one copy to ensure against loss or damage. Punch cards are usually the first medium used to record the data. These should be kept in a dry condition at cool room temperature so as to maintain their flatness and correct size necessary for use with machinery. The storage of cards is facilitated by the use of special metal drawers which also provide a space for clear labelling of the data. If cards are to be the medium of input into the tabulating machinery, it is advisable to keep an extra backup copy so that damaged cards can be readily reproduced. In many cases, however, the data are read from the punch cards onto magnetic tape or disk. Where data files are large and card input is cumbersome and slow, a backup copy on tape or disk may be more useful than punch card backup.

Whenever data are copied from one medium to another or when two copies are made in the same medium, at least part of the new copy should be printed out so that it can be checked against the original. At the same time totals and subtotals should be checked to ensure that all the data have been copied. Each new set of records should be fully documented, that is, a written description of the location and format of the data together with other pertinent information should be prepared. A copy of the code book should be included as part of the documentation as well as dataset labels and control totals.

C. ACCESSIBILITY AND CONFIDENTIALITY

If the dataset is recorded on tape or disk its accessibility should be restricted to those who are legitimately entitled to use it. This is particularly important where the data are of a confidential nature or where confidentiality of information has been promised to the respondents. Making the data available to other individuals or agencies may result in use of the data in ways other than the original purpose, which may be harmful or embarrassing to the respondents. Since such an occurrence would seriously impair the credibility of the researcher as well as reduce the cooperation of respondents in future surveys, accessibility should be tightly controlled.

VI. Planning for Tabulations

A. INVENTORYING DATA PROCESSING EQUIPMENT

As previously noted, it will be necessary to have a thorough knowledge of the capabilities and limitations of the equipment to be used at an early stage of planning. This section will therefore briefly discuss some of the most commonly encountered types of data processing equipment in relation to their use in preparing tabulations.

1. Sorters and Counter-Sorters

Sorters are machines with electric or electronic sensing devices which can read the punches in any of the 80 columns in a standard card and route the card to a pocket or bin corresponding to the punch. Most sorters count the number of cards read, while some are equipped to count the number of cards falling into each pocket. Sorters with the latter type of counting equipment are termed *counter-sorters*.

If a variable has 10 or fewer categories and hence occupies a single column, one pass through the counter-sorter will produce a frequency distribution for the variable. Where codes require more than one column, the cards need to be sorted once for each column. For example, to obtain a frequency distribution for a two-digit code, the deck of cards is first sorted on the column holding the tens position punch. This produces a set of subdecks, 0, 1, 2 . . . etc., each of which is then sorted on the units columns to obtain the full detail of the code.

The logic of producing cross-tabulations on a counter-sorter is similar to that of producing frequency distributions for codes extending over more than one column. The card deck is sorted into subdecks according to the codes assigned to the first variable; each subdeck is then further sorted into the categories of the second variable.

So long as all the variables of interest pertaining to a given unit of analysis are confined to a single card, the counter-sorter can produce any cross-classification desired. Its limitations are many, however, and may be summarized as follows:

a) Although counter-sorters vary greatly with respect to their processing speed, they are in general slower than alternative *mechanical* data processing devices such as tabulators.
b) Most counter-sorters can read only one

column at a time, necessitating the creation of subdecks; complex analyses can thus become quite cumbersome.

c) Each time a deck of cards is sorted, the risk of machine jams and damage to cards increases; substantial amounts of time can be lost in unjamming the machine and reproducing damaged cards.

d) Data from the card counter must be manually transcribed onto tabulation forms; this introduces the risk of human error.

e) The data produced are in the form of simple counts; further analyses (percentage distributions, statistical tests, etc.) must be carried out separately.

2. *Collators and Reproducers*

When the data for a single unit of analysis (a household or an individual, for example) extend over more than one punch card, *collators* and *reproducers* may be useful. The circuit board which governs the operation of the collator can be wired to read and compare the characters in several columns of the punch card. Either one or two decks can be read simultaneously. If two decks have been sorted into a given sequence, on the basis of household number, for example, the collator can be wired to merge the two decks so that punch cards corresponding to the same individual or household are put in order.

When there are two or more records per case, reproducers can be used to create new analysis decks. Thus, if a cross-tabulation of variables appearing on different cards is desired, the reproducer can be used to create a new punch card deck in which variables appear on the same card.

3. *Tabulators*

If electronic computers are not available, access to a *tabulator* can lead to considerable savings in time and effort over preparing tabulations on a counter-sorter. Originally developed for accounting and billing purposes, tabulators can be wired to produce relatively complex statistical reports. The basic functions of tabulators include reading, selecting, adding, subtracting, counting, printing, and sometimes punching. These activities are usually governed by the wiring of circuit boards. Depending on how the user has wired the circuit board, the machine can read the data for one or more variables from each punch card. Tabulators can be used to produce frequency distributions, cross-tabulations, and sums of variables. Tabulators have three major advantages over counter-sorters:

a) Although cards must be sorted into the categories for which counts or totals are desired, they need not be further sorted by each characteristic to be tabulated. Thus, if age distributions by sex and marital status are desired, and the deck is sorted by sex and marital status (but not by age) the machine can be wired to read age and to count the number of records read in each age category. It will print the age distribution for each sex-marital status category as well as a total for each sex and a grand total for all records run.

b) The addition and subtraction capabilities of the tabulator makes it possible to carry out certain arithmetic operations on the tabulator. If mean number of children per woman is desired, the machine can simultaneously count the number of women and accumulate the number of children, thus providing numerator and denominator data for the required calculation.

c) Counts and totals are printed automatically by the tabulator, thereby avoiding errors of transcription and illegible entries. Formats are fairly flexible thus raising the possibility of producing final tables by photocopy. Machine error in printing, however, remains a possibility.

On the other hand, tabulators are not, in general, capable of multiplication or division and hence cannot directly produce the statistics generally needed for analysis. Further, though most operations can be carried out more quickly on a tabulator than on a sorter, they are slow relative to electronic computers.

4. *Computers*

There are numerous different computer models varying in size, complexity, and sophistication. The size of a computer is measured by memory space which determines the amount of data that can be processed and the size of the programs that can be used. Auxiliary equipment in the form of input and output devices linked to the computer are also required. Thus, data may be read in from a card reader, tape drive, or disk, while the output may be in the form of line printing, card punching, graph plotting, and writing onto tape or disk.

The usual method of submitting a job to the computer is to feed a file or deck of instruction cards through the card reader. This file consists of three sets of instructions:

a) Control cards that give pertinent instructions to the computer. These include information on where the data are stored, which devices are to be used for input and output, how much computer memory is needed, which language has been used in writing the program, and other requirements for completing the job.

b) Program cards that instruct the computer to process the data according to the analysis required. If the control cards have indicated that a packaged program is to be used, the program cards may consist of instruction cards specific to that program.

c) Data cards that contain the data to be processed. These cards will not be present if the data file is on tape or disk.

The output resulting from the instructions fed to the computer is written out by the printer. This output should consist of the calculations intended by the researcher. Often, however, an instruction will have been improperly stated or punched. In such cases, the job output will indicate an execution failure and give information on the nature of the error.

This procedure of reading in an entire job at one time and obtaining output based on a complete set of instructions is called *batch processing*. Batch jobs may be submitted via any card reader linked to a computer including card readers located at some distance away from it. Output may likewise be written at any printer linked to the system.

In batch processing, the user effectively loses control of the job from the time it is submitted to the time it is returned on the output device selected. At some computer installations, an alternative type of processing, *interactive processing*, is available. Under interactive processing, jobs are entered through a terminal. This piece of equipment looks rather like a typewriter except that the keyboard may have some unfamiliar characters. Connection to the system is made by dialing the system telephone number; communication after connection is via the keyboard. Communications from the computer to the user are printed out automatically. The user can instruct the computer step by step on the basis of the outcome of previous steps. Interactive terminals may also give users access to "conversational" programs where the user merely answers a series of questions and enters the data at the appropriate time.

In addition to interactive operations, terminals may be used to submit and retrieve batch jobs. In such applications, typewritten instructions replace punch cards, though their content is identical.

B. INVENTORYING SOFTWARE EQUIPMENT

1. Packaged Programs

Numerous computer packaged programs have been devised to deal with the processing of data. Some perform only the basic frequency counts and cross-tabulations while others are able to do more complex analyses. Each package has its own simple "language" of commands which merely requires the user to give basic information about the data and the required analysis. It is important, however, that the user understands what each program does and that the data are suited to the analysis.

Use of a packaged program should be decided upon at the planning stage of the survey (as indicated in the PERT chart in Figure 1) so that the coding and punching can be designed to be compatible with the package requirements. If the package is not already installed in a convenient computing system, the feasibility of installing it should be looked into very early in the planning process. Some computers are simply too small to accommodate the larger packaged programs, and even if this is no problem the installation and debugging period may be longer than time will allow.

2. Utility Programs

Various utility programs are usually available at each computer center. These provide the means of manipulating data into the form required for use. Copying programs falls into this category; data can be copied from and to cards, tape, or disk and can also be printed out on paper. Other utility programs provide a means for adding to or removing from a dataset, for carrying out basic editing routines, and for transforming codes.

3. Writing Programs

Unless one wants to do nonstandard analysis or unless packaged programs are unavailable there is little need to know how to program, though some knowledge of the art may facilitate the use of packaged and utility programs. Computer programs are written in computer "languages" such as FORTRAN, ALGOL, COBOL, BASIC, ASSEMBLY, and PL/1 which are translated by the compiler into machine instructions. Not all computer systems can compile all languages. Hence, it is important to know which languages can be translated before beginning to write programs.

VII. Tabulation and Analysis

A. USES OF SIMPLE FREQUENCY COUNTS AND CROSS-TABULATIONS

Most packaged programs contain routines for producing frequency distributions (numbers of cases falling into each coding category) and cross-tabulations (which show frequency distributions for one variable cross-classified by the categories of one or more other variables). See Figure 5. If such programs are not available, the same information can be obtained through use of sorters or tabulators or through individually written programs.

Information from simple frequency counts is not only of interest in its own right, but is also often used in transforming variables, keeping checks on category totals, and refining plans for further analysis. The first and most basic check is to be certain that the count of records is equal to the number of records in each batch, as indicated on the batch control chart. If discrepancies in these totals and subtotals are found, immediate steps should be taken to find and rectify the errors. This may involve elimination of duplicate records as well as location of missing ones.

The next step is to determine the number of cases per category for each variable. These frequencies serve as important control totals throughout the analysis. Where variables are cross-tabulated the row and column totals should give the distributions of the two variables, and when subgroups are analyzed the number of cases accounted for in each table should equal the relevant subgroup total. Such controls guard against the possibilities of omitting part of the data from the analysis and of including data that should be omitted.

As noted earlier, the categories used for coding usually show more detail than is needed for final tabulations. Simple frequency distributions and cross-tabulations provide a useful guide for transforming detailed codes into fewer, broader categories. Thus if the distribution of a variable is such that some categories contain very few cases, it is advisable to combine adjacent or meaningful groups of categories for use in final tabulations. For example, if educational attainment has been coded in terms of number of years of school completed, one possibility might be to group the data into two categories, one for those whose highest year completed was in primary school and one for those completing one or more years of secondary school.

Transformation of variables into fewer categories may also need to take the relations between two or more variables into account. This is most easily done by use of cross-tabulations showing the full detail for each variable. Coding categories can then be constructed to ensure that the number of cases in each category combination is not so small that the results are meaningless or that confidentiality is violated.

B. THE DESIGN OF FINAL TABULATIONS

Since final tabulations appear in the research report, considerations of economy and clarity should influence their design. Tabulations showing two variables are relatively straightforward, but those involving three or more variables require careful arrangement. For example, in presenting data on number of children ever born by age of mother, marital status, and rural-urban residence, one must decide whether the basic focus of attention is to be on differences among marital status groups or between rural and urban populations. If marital status comparisons are of chief interest, then data for each marital status group may be shown in panels, one for urban and one for rural; if emphasis is placed on rural-urban differences, the subheadings rural and urban should appear under each marital status category.

The above example leads to another point relevant to table economy. Sometimes categories of relatively marginal interest need not be presented directly, if they can be derived by the reader from the data given. Thus, for example, data on the total number of children ever born by age of women are often presented for all women, for ever-married women, and for currently married women since these are the groups of greatest interest. From such a table it is also possible to obtain data on children ever born to never-married women by subtracting the results for ever-married women from those for all women. Similarly, information on "other" ever-married women can be obtained by subtracting the results for currently married women from those for ever-married women.

Still another consideration is whether the data are most usefully presented as absolute numbers or as relative measures such as percentages or means. In part, the answer to this question is determined by the nature of the report, in part, by the general availability of the data to other potential users. Comparison of absolute numbers is awkward, especially when the populations being compared are of different sizes; on the other hand, if it is expected that many readers will want to carry out additional analyses, inclusion of absolute numbers is warranted. Hence, in most cases, both absolute and relative quantities are needed; often it is convenient to use relative amounts in the main body of the report and to provide absolute numbers in an appendix.

C. WEIGHTING

If the data come from a probability sample in which the units are selected with unequal probabilities, appropriate weighting is required to make the sample representative of the universe. Such weighting may be incorporated in the tabulation program or may be done independently after tabulations are prepared. In either case, care must be taken to ensure that the weighting routine corresponds to that implied by the sample design. Summary statistics, such as those discussed in the next section, should take account of the appropriate weights.

D. SUMMARY STATISTICS

The various statistics that can be calculated from frequency counts and cross-tabulations are briefly reviewed in this section. However, it should be noted that not all statistical measures, including those measures automatically produced by packaged programs, are meaningfully applied to every variable. Thus, for example, means and variances are not applicable to ordinal or nominal scales and rank statistics are inapplicable to nominal scales.

The statistics associated with simple frequency counts are those that describe the distribution of a variable. They include measures of central tendency or location such as the mode, mean or proportion, median, and other percentiles, as well as measures of variation including the standard deviation and variance, range and inter-quartile range.

Those statistics associated with cross-tabulations usually describe the strength of the relationship between two variables. The basic statistic for indicating the presence and degree of dependence between two variables is chi-square. Other related statistics including the contingency coefficient are available. In addition, there are various nonparametric statistics available which indicate strength of association between ranks and nominal scales.

E. MORE COMPLEX ANALYSIS

After the preliminary analysis consisting of frequency counts and cross-tabulations has been completed, more complex relations within the data may be investigated. Although this should have been in mind at the planning stage, its precise nature often cannot be determined until earlier results have been examined. The more complex stages of analysis may include regression, correlation, and analyses of variance and covariance. Multivariate techniques such as multiple regression, principal components, and factor analysis may also be employed.

These more complex techniques are extremely valuable in that they may summarize complicated networks of interrelationships into a more readily comprehensible form. It is important to remember however that many of them are not readily applicable to complex stratified survey designs and are sensitive to extreme values which may distort the results. Hence detailed study of frequency distributions and cross-tabulations is required.

VIII. Costs and Quality Control

Data are rarely, if ever, perfect. When this fact is accepted, it becomes possible to view the problem of data accuracy as one of establishing an acceptable level of accuracy to be achieved at a given cost and ensuring that the data remain within the established quality limits. A considerable literature on methods of controlling quality has already been developed. (See, for example, Fasteau et al., 1964; Nordbotten, 1965; Szameitat and Zindler, 1965; Stuart, 1966; Minton, 1969; O'Reagan, 1969; Simmons, 1972). Although many of these deal with computer applications to quality control, the underlying ideas are often general enough to be applied to other situations.

Editing can be regarded as a kind of quality control for fieldwork. Levels of nonresponse and inconsistencies in the reported information are useful indicators of interviewing difficulties, and if noted in early stages, during the pretest for example, can be of value in improving fieldwork.

Once something is known about the various types of error and their magnitudes, careful thought should be given to priorities and strategies for error reduction. Nordbotten (1965) has distinguished four types of error: rare and small, frequent and small, rare and large, and frequent and large. He suggests that, because the first two types of error are small, they are relatively unimportant. Frequent, large errors are likely to be caught. The most troublesome case, therefore, is that of rare, large errors. The notion of bias could be usefully added to this classification—errors that produce random fluctuations in the data can usually be regarded as less serious than systematic errors. Finally, the cost of error reduction should be considered in establishing error reduction priorities.

Developing a strategy for both error and cost reduction requires thorough knowledge of the uses of the data as well as the total data collection system, including information about the genesis and etiology of errors of various types, and estimates of the cost of improving the quality of the data at each stage of data collection. Given this information the trade-offs between costs and allowable error can be considered. Thus, for example, if small random errors prove expensive to correct, they can perhaps be tolerated. In addition, it is possible to compare and minimize the costs of making corrections at different stages of data processing. Such cost minimization must be tailored to individual situations. Thus, in some systems, quality may be least expensively improved by hiring additional interviewers or interviewing supervisors, while under conditions of high manpower costs, computer edits of defective data may produce the minimum cost for a given level of quality. The essential condition for cost minimization is therefore thorough understanding of the system and of alternative approaches to accomplishing each phase of the work.

* * * * * *

ACKNOWLEDGMENTS

We wish to thank Mary Grace Kovar for her very helpful comments on earlier drafts of this report. Thanks are also due Deborah W. Rice for her patient typing of both the final manuscript and various earlier drafts. The errors and ambiguities remaining are, of course, the responsibility of the authors.

IX. Bibliography

Cantrelle, Pierre. 1974. Systems of Demographic Measurement. Data Collection Systems: La Méthode de l'Observation Démographique Suivie par Enquête à Passages Répétés (OS/EPR). Scientific Report Series No. 14. Chapel Hill, N.C.: International Program of Laboratories for Population Statistics.

Cook, L. 1966. Program Evaluation and Review Technique: Applications in Education. Washington, D.C.: U.S. Government Printing Office.

Crittenden, Kathleen S., and Richard J. Hill. 1971. Coding Reliability and Validity of Interview Data. American Sociological Review 36(6):1073–1080.

Downham, J. S. 1955. The Function of Coding. The Incorporated Statistician. Supplement to Vol. 5(4):73–81.

Fasteau, Herman H., J. Jack Ingram, and George Minton. 1964. Control of Quality of Coding in the 1960 Censuses. Journal of the American Statistical Association 59(305):120–132.

Freund, R. J., and H. O. Hartley. 1967. A Procedure for Automatic Data Editing. Journal of the American Statistical Association 62(318):341–352.

Frisbie, Bruce, and Seymour Sudman. 1968. The Use of Computers in Coding Free Responses. Public Opinion Quarterly 32(1):216–232.

Harris, Amelia I. 1955. The Work of a Coding Section· The Incorporated Statistician. Supplement to Vol. 5(4):82–91.

Kammeyer, Kenneth C. W., and Julius A. Roth. 1971. Coding Responses to Open-Ended Questions. Sociological Methodology, 1971. San Francisco: Jossey-Bass, Inc.

Minton, George. 1969. Inspection and Correction Error in Data Processing. Journal of the American Statistical Association 64(328):1256–1275.

———. 1970. Some Decision Rules for Administrative Applications of Quality Control. Journal of Quality Technology 2(2):86–98.

Moser, C. A., and G. Kalton. 1971. Survey Methods in Social Investigation. London: Heinemann Educational Books Limited.

Muehl, D. 1961. A Manual for Coders. Ann Arbor, Mich.: Survey Research Center, University of Michigan.

Nordbotten, Svein. 1965. The Efficiency of Automatic Detection and Correction of Errors in Individual Observations As Compared with Other Means for Improving the Quality of Statistics. Bulletin of the International Statistical Institute, Proceedings of the 35th Session, 1965. Vol. 41, Book 1:417–441.

Omaboe, E. N., and K. T. deGraft-Johnson. 1967. Possibilities for Evaluation of Census or Survey Data in Developing Countries. Bulletin of the International Statistical Institute, Proceedings of the 36th Session, Sydney, 1967. Volume 42, Book 1:91–100.

O'Reagan, Robert T. 1969. Relative Costs of Computerized Error Inspection Plans. Journal of the American Statistical Association 64(328):1245–1255.

Parten, Mildred. 1966. Surveys, Polls, and Samples: Practical Procedures. New York: Cooper Square Publishers, Inc.

PERT Orientation and Training Center. n.d. PERT Fundamentals. Washington, D.C.

Pritzker, Leon, Jack Ogus, and Morris H. Hansen. 1965. Computer Editing Methods—Some Applications and Results. Bulletin of the International Statistical Institute, Proceedings of the 35th Session, 1965. Vol. 41, Book 1:442–466.

Sessions, Frank Q., Robert J. Epley, and Edward O. Moe. 1966. The Development, Reliability, and Validity of an All-Purpose Optical Scanner Questionnaire Form. Public Opinion Quarterly 30(3):423–428.

Shryock, Henry S., Jacob S. Siegel, and Associates. 1973. The Methods and Materials of Demography. 2nd Printing (Rev). Washington, D.C.: U.S. Government Printing Office.

Simmons, Walt R. 1972. Operational Control of Sample Surveys. Manual Series No. 2. Chapel Hill, N.C.: International Program of Laboratories for Population Statistics.

Simpson, H. R. 1961. The Analysis of Survey Data on an Electronic Computer. Journal of the Royal Statistical Society, Series A, 124:219–226.

Stuart, Walter J. 1966. Computer Editing of Survey Data—Five Years of Experience in BLS Manpower Surveys. Journal of the American Statistical Association 61(314, part 1):375–383.

Sudman, Seymour. 1967. Reducing the Cost of Surveys. Chicago: Aldine Publishing Company.

Szameitat, Klaus, and Hans-Joachim Zindler. 1965. The Reduction of Errors in Statistics by Automatic Corrections. Bulletin of the International Statistical Institute, Proceedings of the 35th Session, 1965. Vol. 41, Book 1:395–417.

United Nations. 1955. Handbook of Vital Statistics Methods. Studies in Methods, Series F, No. 7. ST/STAT/SER.F/7. New York.

———. Statistical Office, and Food and Agriculture Organization of the United Nations. 1959. Handbook on Data Processing Methods, Part I, Provisional Edition. New York and Rome.

U.S. Bureau of the Census. 1972. Evaluation and Research Program of the U.S. Censuses of Population and Housing 1960. Effects of Coders. Series ER60, No. 9. Washington, D.C.

Woodward, Julian L., and Jack DeLott. 1952. Field Coding Versus Office Coding. Public Opinion Quarterly 16(3):432–436.

Wooldridge, Susan, Colin R. Corder, and Claude R. Johnson. 1973. Security Standards for Data Processing. New York: Wiley.

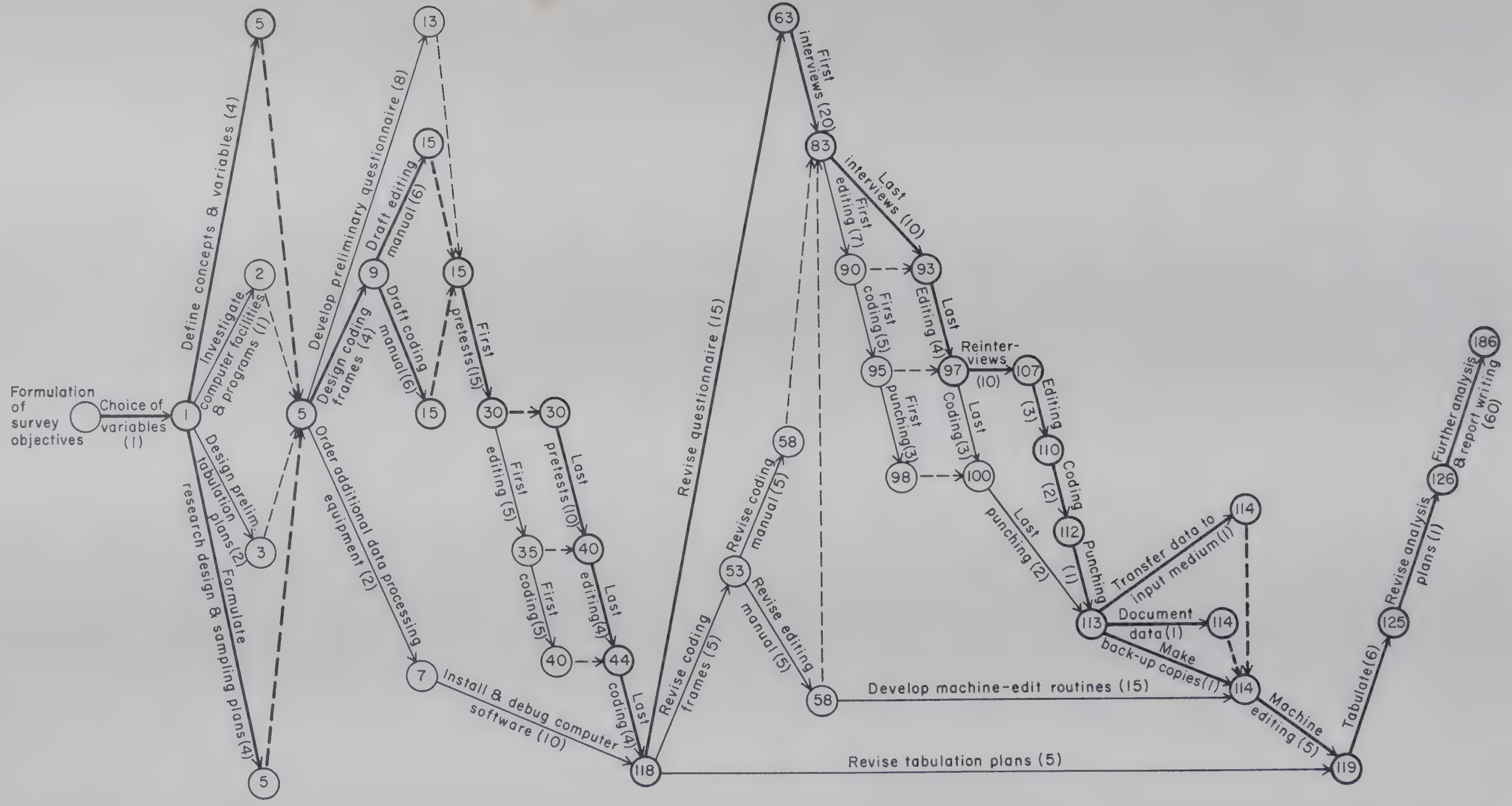

FIGURE 1

PERT Chart for the Survey Process

1) Figures in circles indicate latest starting time in days for the subsequent activity.
2) Figures in parentheses indicate the expected time in days for each activity.
3) Heavy lines indicate the critical path.

FIGURE 2

Example of Coding Frames

CODE BOOK FOR FERTILITY QUESTIONNAIRE

Question No.	Column No.	Variable Number and Code
41	13	53 Approval of contraception 1 Approve [*Code 98 in 14–15*] 2 Disapprove 3 Uncertain 9 No response
41a & b	14–15	54 Conditions when all right Yes to 41a [*Code first response only*]: 01 To protect the health of the mother 02 If there is serious danger of the child's being defective 03 So that children already born can be given more care and attention 04 If the family is poor or cannot afford more children 05 In order to provide for the education of children already born 06 If husband and wife are not getting along 07 If the family already has the number of children it wants 08 So both husband and wife can work 09 To maintain or improve family welfare or happiness 10 So husband and wife are free to do other things 11 Because a smaller population would be good for the country 12 Other 97 No to 41a 98 Not applicable 99 No response
42	. . .	. . .
43	. . .	. . .
44	. . .	. . .
45	27–30	59 Methods known [*Code the sum of all circled responses*] 9998 If "no" to Q 45 9999 No response
46a	31–34	60 Methods ever used [*Code the sum of all circled responses*] 9998 If "no" to Q 46 9999 No response
47	35–36	61 Method used most recently [*Code circled response*] 98 If "no" to either Q 45 or 46 99 No response

FIGURE 3

Part of Completed and Coded Questionnaire Used in Interview of Ever-Married Women

Column No.

41. Many couples do something to delay or prevent a pregnancy, so that they can have just the number of children that they want, and have them when they want them. How do you feel about this? Would you say that you approve, disapprove, or feel uncertain about this?

1. Approve *Skip to Q 42*
2. Disapprove — 13
[✓ 3.] Uncertain

If Disapproves or Uncertain, ask:

41(a) Do you think there are any conditions under which it is all right for married couples to do something to delay or prevent a pregnancy?

[✓ Yes] No *Skip to Q 42*

If Yes, Ask:

41(b) What are those conditions? (Are there any others?)

Enter here: – if mothers health in danger — 14–15 0
– if family too poor to
support another child

42. . . .

43. . . .

44. . . .

45. Do you know any methods that are used by married couples to delay or prevent a pregnancy?

[✓ Yes] No *Code 9998 in cols 27–30; 9998 in cols 31–34; 98 in cols 35–36.*

If Yes, Ask: *Skip to Q 48*

45(a) What methods have you heard about?

Circle the appropriate codes in column 1 of list in table below.

46. Have you ever used any of these methods?

[✓ Yes] No *Code 9998 in cols 31–34; 98 in cols 35–36.*

If Yes, Ask: *Skip to Q 48*

46(a) Which methods have you ever used?

Circle the appropriate codes in column 2 of list in table below.

47. What method have you used most recently?

Circle the appropriate code in column 3.

Figure 3 (*Continued*)

	Method	Answer to	Answer to	Answer to
		Q 45 (a)	Q 46	Q 47
		(1)	(2)	(3)
A	Abstinence or living apart	1	1	01
B	Rhythm (Safe period)	2	2	02
C	Withdrawal	[✓ 4]	[✓ 4]	03
D	Douche	8	8	04
E	Breast feeding	16	16	05
F	Condom	[✓ 32]	[✓ 32]	06
G	Diaphragm	64	64	07
H	Foam, jelly or cream, suppositories	128	128	08
I	IUD (Loop)	[✓ 256]	[✓ 256]	[✓ 09]
J	Pill	[✓ 512]	512	10
K	Male sterilization (Vasectomy)	[✓ 1024]	1024	11
L	Female sterilization (Tubal ligation)	2048	2048	12
M	Other *Enter here:* ____________	4096	4096	13
	Sum	1828	0292	09
	Column Numbers	27–30	31–34	35–36

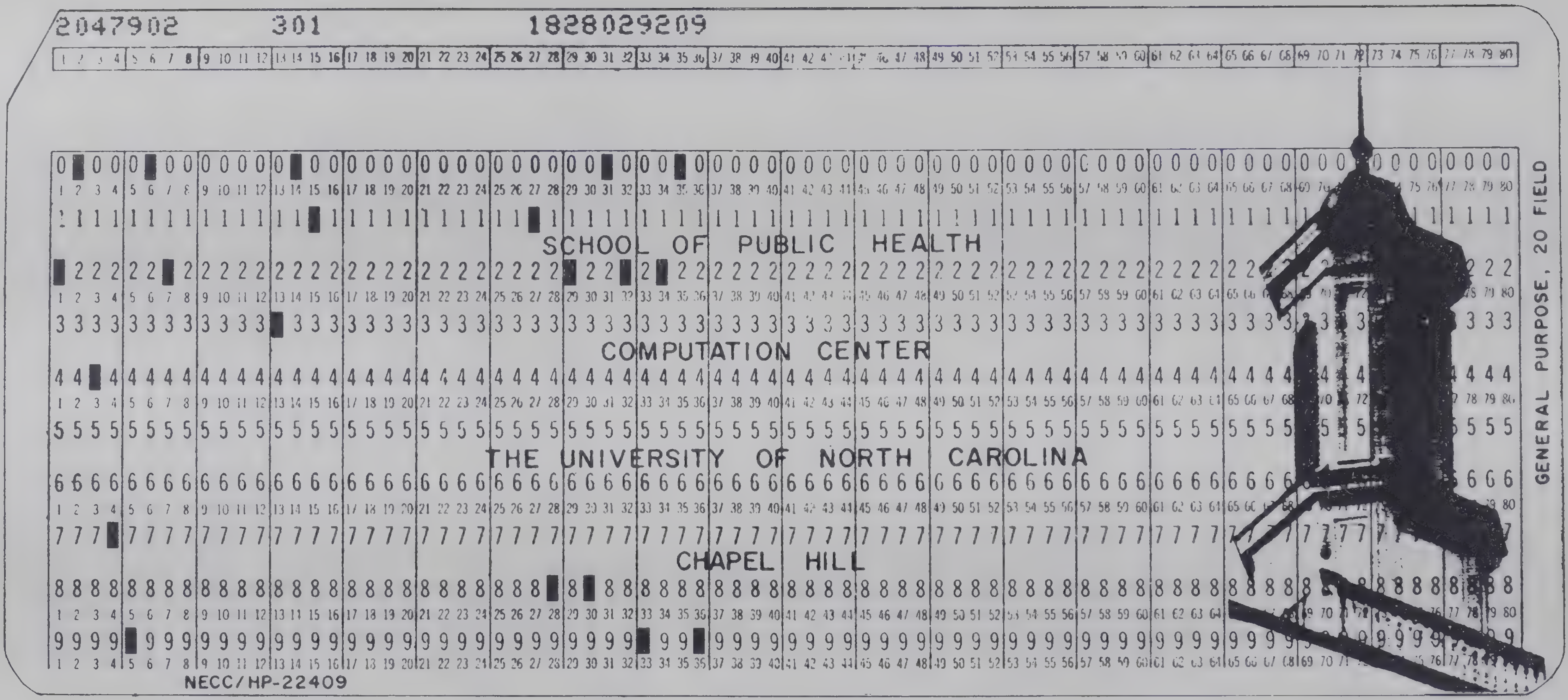

FIGURE 4

Punch Card Containing Data Corresponding to the Coded Questionnaire in Figure 3

The first seven columns contain the following identification:

col 1	*Card number*
cols 2–5	*Household number*
cols 6–7	*Individual number within household*

Figure 5

Frequency Distributions and Cross-Tabulations Relevant to Questions Shown in Figure 3

5-A. Frequency Distribution of Variable 53 Including Nonresponse

			Cumulative	
	%	Number	%	Number
Approve	61.2	506	61.2	506
Disapprove	10.9	90	72.1	596
Uncertain	23.8	197	95.9	793
No response	4.1	34	100.0	827
Total	100.0	827		

5-B. Frequency Distribution of Variable 53 After Imputation of Missing Values

			Cumulative	
	%	Number	%	Number
Approve	63.7	527	63.7	527
Disapprove	11.4	94	75.1	621
Uncertain	24.9	206	100.0	827
Total	100.0	827		

5-C. Frequency Distribution of Variable 60: An Example of How to Obtain Simple Frequency Counts from Geometric Codes

		Method Ever Used												
Code	Frequency	A	B	C	D	E	F	G	H	I	J	K	L	M
1	0	–	–	–	–	–	–	–	–	–	–	–	–	–
2	0	–	–	–	–	–	–	–	–	–	–	–	–	–
3	1	1	1	–	–	–	–	–	–	–	–	–	–	–
4	6	–	–	6	–	–	–	–	–	–	–	–	–	–
5	8	8	–	8	–	–	–	–	–	–	–	–	–	–
6	11	–	11	11	–	–	–	–	–	–	–	–	–	–
7	20	20	20	20	–	–	–	–	–	–	–	–	–	–
8190														
8191														
Total	797*	430	458	534	258	570	390	335	329	374	393	65	34	9
%	100.0	54.0	57.5	67.0	32.4	71.5	49.0	42.0	41.3	47.0	49.3	8.2	4.3	1.1

* *Since 30 respondents replied "No" to question 46, the total number of responses to question 46(a) is 827 − 30 = 797.*

FIGURE 5 (*cont.*)

5-D. CROSS-TABULATION OF VARIABLE 53 BY VARIABLE 60

	Method Ever Used	Approve	Disapprove	Uncertain	Total
A	Abstinence or living apart	244	55	131	430
B	Rhythm (Safe period)	356	25	77	458
C	Withdrawal	301	51	182	534
D	Douche	226	2	32	260
E	Breast feeding	384	42	144	570
F	Condom	297	0	93	390
G	Diaphragm	321	0	14	335
H	Foam, jelly or cream, suppositories	234	0	95	329
I	IUD (Loop)	352	0	22	374
J	Pill	370	0	23	393
K	Male sterilization (Vasectomy)	65	0	0	65
L	Female sterilization (Tubal ligation)	33	0	1	34
M	Other	7	0	2	9
	Number of responses*	527	69	201	797

* *These are less than the frequency counts in Figure 5-B since 30 respondents reported never having used a contraceptive method.*

5-E. CROSS-TABULATION OF VARIABLE 53 BY VARIABLE 61

	Method Used Most Recently	Approve	Disapprove	Uncertain	Total
A	Abstinence or living apart	0	18	2	20
B	Rhythm (Safe period)	1	14	1	16
C	Withdrawal	1	9	29	39
D	Douche	18	1	24	43
E	Breast feeding	6	27	23	56
F	Condom	101	0	36	137
G	Diaphragm	26	0	10	36
H	Foam, jelly or cream, suppositories	16	0	37	53
I	IUD (Loop)	121	0	20	141
J	Pill	139	0	18	157
K	Male sterilization (Vasectomy)	65	0	0	65
L	Female sterilization (Tubal ligation)	33	0	1	34
M	Other	0	0	0	0
	Total*	527	69	201	797

* *See footnote to Figure 5-D.*

Appendix: Glossary of Computing Terms

Assemble. Automatic translation of a program written in symbolic language into machine language. Not used when program written in FORTRAN, COBOL, ALGOL, etc. *See also* Compile.

Assembler. Program designed to convert symbolic instructions into a form suitable for execution by the computer. Provides error messages for the programmer to correct the program. *See also* Compiler.

Batch processing. Use of computer system by submission of entire jobs to be loaded into memory and processed sequentially.

Canned program. *See* packaged program.

Card. Information medium used as input to most computers.

Card hopper. Device that holds cards and makes them available to the card-feed mechanism.

Card puncher. Output device that punches out information onto card.

Card reader. Device that reads information from card for input to the computer.

Card stacker. Output device that accumulates punched cards in a deck.

Computer system. Set of related hardware that may be referred to in a general or specific sense:

1) general system is related hardware available from computer manufacturer.
2) specific system is hardware units contiguously installed that operate simultaneously and that are logically connected. Multiple processors may form one system if they are linked, but two unconnected computers form two distinct systems.

Compile. To prepare a machine-language program from a program written in a higher programming language, such as FORTRAN, COBOL, ALGOL, etc. Usually generates more than one machine instruction for each symbolic statement. The process is carried out by a computer program called a compiler.

Compiler. Program to translate a higher language into machine language.

Console. Interface or communication device between operator and computer. May consist of lights, switches, knobs, typewriter, or keyboard.

Control cards or statements. *See* job control.

Conversational time sharing. Use of computer system simultaneously by multiple uses each operating a remote terminal. Conversational refers to the mode in which user and computer cooperate on a give-and-take basis and communicate in a question and answer fashion.

Debug. To test for, locate, and remove mistakes from a program or malfunctions from a computer.

Deck. Collection of punched cards.

Disk. External auxiliary storing device consisting of a thin circular rigid plate rotating very rapidly. One or both of the surfaces are capable of being magnetized. Information is recorded on the surface by creating at a specific spot a magnetic field by means of a recording head. Information is read from disk by the recording head's sensing the magnetized areas.

Drive. *See* Tape drive.

Execute. Carry out instructions in program deck.

File. A collection of related records treated as a unit.

Hardware. Physical, tangible, and permanent components of a computer.

Interactive processing. Use of a computer system from a remote terminal in a conversational mode such that the user has control over each stage of the job.

Interleaving. System under which computer processes small parts of many jobs in turn

Job. Unit of work to be done by a computer. Begins with a job signal and ends with an end-of-job signal.

Job control. Statements that direct the operating system in its functioning.

Keypunch. Keyboard-activated device that punches holes in a card. Vary in size and complexity from portable, manually operated, single-hole-at-a-time punches to semiautomatic (which can skip over columns, duplicate and right-justify numbers within a predefined field that is programmed on a drum card). Characters may also be printed along the top of the card.

Line printer. Device that prints an entire print line in a single operation. Operates at 600 to 1100 lines per minute.

Magnetic tape. Ribbonlike material used to store data in lengthwise sequential position.

Off-Line. Equipment not connected to or temporarily disconnected from the computer.

On-Line. Equipment connected to the computer.

Operating System. Interrelated series of programs and routines designed to make the computer system more efficient and easier to program. Supervises all input and output, job-to-job transitions, provides assemblers, compilers, and program loaders.

Packaged Program. Program that has already been compiled or installed in the operating system and that can be accessed by relatively simple instructions.

Paper tape. *See* punched tape.

Plotter. Output device used to plot graphs.

Punch card. *See* card.

Punched tape. Paper or plastic ribbon having a longitudinal row of small feeder holes, and either five, seven, or eight rows of larger holes for information.

Puncher. *See* card puncher.

Queue. Line or ordered group of jobs waiting to be executed.

Remote console. *See* remote terminal.

Remote job entry. Submission of batch-type jobs from a remote terminal.

Remote terminal. Site at which data can leave or enter the system. Also device or console for entering and receiving data at end of a transmission path. Consists of one input or one output device at least. Increasingly it is used to refer to either a teletypewriter or a cathode-ray-display screen.

Run. Single complete execution of computer program—a single access to a computer.

Software. Intangible components of system including the operating system.

System. *See* computer system, operating system.

Tape. *See* magnetic tape.

Tape drive. Mechanism on which magnetic tape is mounted so that information can be read or recorded. Consists of a supply reel and drive motor, a capstan to move the tape, record- and readback-head, takeup reel, and drive motor.

Terminal. *See* remote terminal.

Time share. Perform several independent processes almost simultaneously on a single high-speed computer. System is interleaving.

Time sharing. Simultaneous utilization of computer system from multiple terminals.

User. Person who uses computer system by submitting a job.

Abstract

Data Processing and Analysis in Demographic Surveys

The processing and analysis phase is crucial in any data collection effort since it bridges the gap between carrying out the fieldwork and making the results of research available to users. Planning for processing and analyses, however, often receives relatively little attention, particularly in the early phases of a research program.

The present manual is intended for those who plan and direct demographic surveys. It focuses on the problem of ensuring that the processing, tabulation, and analysis phases are given adequate attention in the formation of plans for a research program and in the implementation of those plans. Particular emphasis is given to the need for planning tabulation and analysis early in the research program and to the interplay between planning the processing and analysis phases and other aspects of survey planning.

Topics covered include a discussion of the considerations underlying the development of coding frames and code books and the utility of precoded questionnaires. The sequence of processing steps from editing, through coding and punching, to data storage is described in some detail. Examples of applications to demographic surveys are provided. Alternative methods of carrying out each of these steps are discussed in terms of the possible trade-offs between quality control, costs, and efficiency. Considerable emphasis is given to the treatment of nonresponses and methods for imputation when data are missing. Administrative considerations, including the training and supervision of data processing personnel, are also briefly treated.

Guidelines for the planning of tabulations and analysis are also presented. These include the need for careful appraisal of the available software and hardware for preparing preliminary and final tabulations. A glossary of the more frequently used computer terms is included as an aid for those who may be unfamiliar with a field that has rapidly expanded both technically and linguistically. Methods for data reduction and analysis are also discussed, primarily with a view to cautioning against the use of inappropriate techniques. A final section reemphasizes quality and cost control considerations.

Although this manual is primarily intended as an aid for those engaged in the planning of research, enough concrete detail is provided to make it a useful reference for persons directly engaged in the processing and analysis of data.

Resumen

Procesamiento y Análisis de Datos en Encuestas Demográficas

La etapa de procesamiento y análisis de datos es crucial en todo esfuerzo por recolectar datos, dado que es el puente de unión entre el trabajo de campo y la presentación de los resultados de la investigación a los interesados. Sin embargo, generalmente se da poca atención a la planificación del procesamiento y análisis de datos, especialmente durante las primeras etapas de un programa de investigación.

Este manual está dirigido a aquellos que planifican y dirigen encuestas demográficas. Se concentra en el problema de asegurar que las etapas de procesamiento, tabulación y análisis de datos reciban la debida atención en la planificación e implementación de un programa de investigación. Se pone especial énfasis en la necesidad de planificar las tabulaciones y análisis al comienzo de la investigación, y en la interacción entre la planificación de las etapas de procesamiento y análisis y otros aspectos de la planificación de la encuesta.

Los temas cubiertos incluyen una discusión de los aspectos que deben ser considerados en el desarrollo de patrones y manuales de codificación, y sobre la utilidad de los cuestionarios precodificados. Se describe en detalle la secuencia de etapas en el procesamiento de datos, desde la crítica, codificación y perforación, hasta el almacenaje de datos. Se proporcionan, además, ejemplos de aplicaciones a encuestas demográficas. Se discuten métodos alternativos para llevar a cabo cada una de estas etapas, en términos de posibles compromisos entre control de calidad, costos y eficiencia. Se da considerable énfasis al problema de la falta de respuesta y a los métodos para atribuir respuestas cuando faltan datos. También se discuten brevemente algunas consideraciones administrativas, incluyendo entrenamiento y supervisión del personal relacionado con el procesamiento de datos.

Se presentan también guías para la planificación de tabulaciones y análisis. Estas incluyen la necesidad de examinar cuidadosamente las disponibilidades de computador y programas de computación para preparar las tabulaciones preliminares y las finales. Se incluye, además, un glosario de los términos usados más frecuentemente en computación, como ayuda para aquellas personas que no estén familiarizadas con un campo que ha crecido rápidamente tanto técnica como linguísticamente. Se discuten además métodos para la reducción y análisis de datos, principalmente con el objeto de prevenir el uso de técnicas inadecuadas. La sección final reenfatiza la importancia del control de calidad y de costos.

A pesar de que este manual está dirigido principalmente a aquellas personas que trabajan en la planificación de la investigación, se proporcionan suficientes detalles concretos que lo hacen útil como referencia a personas que trabajan directamente en el procesamiento y análisis de datos.

Résumé

Dépouillement et analyse dans les enquêtes démographiques

Dans tout effort de collecte de données, la phase du dépouillement et de l'analyse qui comble le fossé entre le travail sur le terrain et la mise à la disposition des utilisateurs des résultats de la recherche, constitue une étape cruciale. Cependant, les plans de dépouillement et d'analyse reçoivent bien souvent relativement peu d'attention en particulier dans les premiers temps d'un programme de recherche.

Cette publication s'adresse à ceux qui organisent et dirigent des enquêtes démographiques. Elle insiste sur la nécessité d'accorder une attention suffisante aux étapes de dépouillement, d'élaboration des tableaux et d'analyse au moment de la planification d'un programme de recherche et de sa mise en oeuvre. On insiste particulièrement sur la nécessité de prévoir, tôt dans le programme de recherche, l'élaboration de tableaux et l'analyse, ainsi que sur les relations réciproques entre les étapes de dépouillement et d'analyse et les autres aspects de l'organisation de l'enquête.

Les sujets abordés comprennent une discussion des considérations sous-jacentes à l'élaboration de systèmes de codages, des livres de codes et de l'utilité des questionnaires précodés. La suite des opérations depuis le contrôle, le codage et la perforation jusqu'au stockage des données est décrite en détail; on donne des exemples d'application à des enquêtes démographiques. Des méthodes différentes sont exposées pour chacune de ces étapes en tenant compte des alternatives entre contrôle de qualité, coût et efficacité. On donne une grande importance au traitement des non réponses et aux méthodes d'attribution quand les données manquent. Des considérations administratives y compris la formation et le contrôle du personnel chargé du traitement des données sont brièvement traitées.

Des guides pour la prévision des tableaux et l'analyse sont également présentés. Ceci comprend la nécessité d'une soigneuse appréciation des disponibilités en terme d'ordinateur et de programmes pour l'élaboration préliminaire et finale des tableaux. Un glossaire des termes courants d'ordinateur est destiné à aider ceux qui peuvent ne pas être familiarisés avec ce domaine qui s'est rapidement développé tant au niveau technique qu'au niveau du vocabulaire. Des méthodes de statistiques descriptives et analytiques sont aussi présentées dans le but essentiel de mettre en garde contre les techniques inappropriées. La dernière section insiste à nouveau sur les considérations de contrôle de qualité et coût.

Bien que ce manuel s'adresse en premier lieu à ceux qui sont engagés dans la prévision et l'organisation de la recherche, il comprend suffisamment de détails pratiques pour en faire une référence utile pour ceux qui sont directement impliqués dans le traitement et l'analyse des données.

The Manual Series (Cont.)

4. PLANNING THE RESEARCH INTERVIEW by John Scott and Eliska Chanlett

5. THE PREPARATION OF AN INVENTORY OF DEMOGRAPHIC DATA FOR SOCIAL AND ECONOMIC PLANNING by Richard E. Bilsborrow

6. DATA PROCESSING AND ANALYSIS IN DEMOGRAPHIC SURVEYS by Heather Booth and Joan W. Lingner

The Reprint Series

1. ON A METHOD OF ESTIMATING BIRTH AND DEATH RATES AND THE EXTENT OF REGISTRATION by C. Chandra Sekar and W. Edwards Deming

2. THE USE OF SAMPLING FOR VITAL REGISTRATION AND VITAL STATISTICS by Philip M. Hauser

3. THE DESIGN OF AN EXPERIMENTAL PROCEDURE FOR OBTAINING ACCURATE VITAL STATISTICS by Ansley J. Coale *and*

   SOME RESULTS FROM ASIAN POPULATION GROWTH STUDIES by William Seltzer

4. A CRITIQUE OF METHODS FOR ESTIMATING POPULATION GROWTH IN COUNTRIES WITH LIMITED DATA by W. Brass

5. FIELD EXPERIENCE IN ESTIMATING POPULATION GROWTH by Patience Lauriat

6. MEASUREMENT OF POPULATION CONTROL PROGRAMS: DESIGN PROBLEMS OF SAMPLE REGISTRATION SYSTEMS by Forrest E. Linder

7. SAMPLE VITAL REGISTRATION EXPERIMENT by Joseph A. Cavanaugh

8. SURVEY METHODS, BASED ON PERIODICALLY REPEATED INTERVIEWS, AIMED AT DETERMINING DEMOGRAPHIC RATES by Carmen Arretx and Jorge L. Somoza

9. EVALUATION OF BIRTH STATISTICS DERIVED RETROSPECTIVELY FROM FERTILITY HISTORIES REPORTED IN A NATIONAL POPULATION SURVEY: UNITED STATES, 1945–1964 by Monroe G. Sirken and Georges Sabagh

10. A COMPARISON OF DIFFERENT SURVEY TECHNIQUES FOR OBTAINING VITAL DATA IN A DEVELOPING COUNTRY by Georges Sabagh and Christopher Scott

11. TECHNICAL PROBLEMS OF MULTIROUND DEMOGRAPHIC SURVEYS by Christopher Scott

12. VITAL EVENT NUMERATION SYSTEM AS A NEW TOOL FOR MEASURING POPULATION CHANGE by Forrest E. Linder

13. PROBLEMS IN DESIGNING INTERVIEW SURVEYS TO MEASURE POPULATION GROWTH by Daniel G. Horvitz

14. ON THE EFFECT OF ERRORS IN THE APPLICATION ON THE CHANDRASEKAR-DEMING TECHNIQUE by William Seltzer and Arjun Adlakha

15. ESTIMACION DEL SUBEMPLEO EN ECUADOR by Centro de Análisis Demográfico

16. FERTILITY ESTIMATES DERIVED FROM INFORMATION ON CHILDREN EVER BORN USING DATA FROM SUCCESSIVE CENSUSES by Carmen Arretx

Occasional Publications

RESEARCH TOPICS FOR THE MEASUREMENT OF POPULATION CHANGE. A CATALOGUE OF STUDY PROTOCOLS by Anders S. Lunde (June 1974)

A GLOSSARY OF SELECTED DEMOGRAPHIC TERMS (July 1974)

A HANDBOOK FOR POPULATION ANALYSTS by Joan W. Lingner (August 1974)

Publications of the POPLAB Program:

The Scientific Report Series

1. THE CONCEPT AND THE PROGRAM OF THE LABORATORIES FOR POPULATION STATISTICS by Forrest E. Linder
2. ORGANIZATION AND METHODS OF THE DUAL REPORT SYSTEM IN TURKEY by Eliska Chanlett
3. SUBJECT MATTER COVERAGE IN THE DUAL-REPORT SYSTEMS OF INDIA, PAKISTAN, TURKEY, AND LIBERIA by James R. Abernathy and Anders S. Lunde
4. THE FIRST AND SECOND INTERNATIONAL POPLAB CONFERENCES
5. ORGANIZATION AND METHODS OF THE DUAL-REPORT SYSTEM IN LIBERIA by Joan W. Lingner
6. ORGANIZATION AND METHODS OF THE DUAL-REPORT SYSTEM IN PAKISTAN by James R. Abernathy
7. THE THIRD ANNUAL INTERNATIONAL POPLAB CONFERENCE
8. THE MINDANAO CENTER FOR POPULATION STUDIES, A PHILIPPINE POPLAB REPORT by Francis C. Madigan
9. ORGANIZATION AND METHODS OF THE DUAL RECORD SYSTEM IN INDIA by Joan W. Lingner and H. Bradley Wells
10. MAPPING FOR RECURRENT RESEARCH: A PHILIPPINE POPLAB REPORT by Francis C. Madigan, et al.
11. FIELD OPERATIONS OF DUAL RECORD TESTS IN THE CIMED POPLAB by CIMED/POPLAB Staff
12. LES DONNEES ET LA RECHERCHE DEMOGRAPHIQUES AU MAROC by Mohamed Rachidi
13. PROCEEDINGS OF THE FOURTH INTERNATIONAL POPLAB CONFERENCE
14. DATA COLLECTION SYSTEMS: LA METHODE DE L'OBSERVATION DEMOGRAPHIQUE SUIVIE PAR ENQUETE A PASSAGES REPETES (OS/EPR) by Pierre Cantrelle
15. DATA COLLECTION SYSTEMS: NATIONAL DUAL RECORD AND RELATED SYSTEMS by H. Bradley Wells
16. SOME ASPECTS OF THE SURVEY DESIGN FOR THE MOROCCAN DUAL RECORD EXPERIMENTAL STUDY by Ivan P. Fellegi
17. PROCEEDINGS OF THE SEMINAR ON THE DESIGN OF THE TURKISH DEMOGRAPHIC SURVEY, 20-22 FEBRUARY 1974, ANKARA; PAPERS PRESENTED, edited by Arjun Adlakha
18. DATA COLLECTION SYSTEMS: EL METODO DE ENCUESTA DEMOGRAFICA DE CELADE by Jorge L. Somoza
19. THE PHILIPPINE POPLAB WITH PARTICULAR REFERENCE TO FIELD OPERATIONS by Francis C. Madigan and James R. Abernathy
20. DATA COLLECTION SYSTEMS: THE CONVENTIONAL VITAL REGISTRATION SYSTEM by Nora P. Powell

The Manual Series

1. MAPPING AND HOUSE NUMBERING by Dorothy S. Cooke
2. OPERATIONAL CONTROL OF SAMPLE SURVEYS by Walt R. Simmons
3. DESIGNING FORMS FOR DEMOGRAPHIC SURVEYS by Monroe G. Sirken

(continued on inside back cover)